Space efficient quantile selection

Input: stream $a_1, \ldots, a_n \in \mathcal{U}$ where $\mathcal{U}$ has order $<$

5 9 3 11 1 4 6 ooo

e.g. numerical data
names w/ alphabetic order
grades

allowed multiple passes

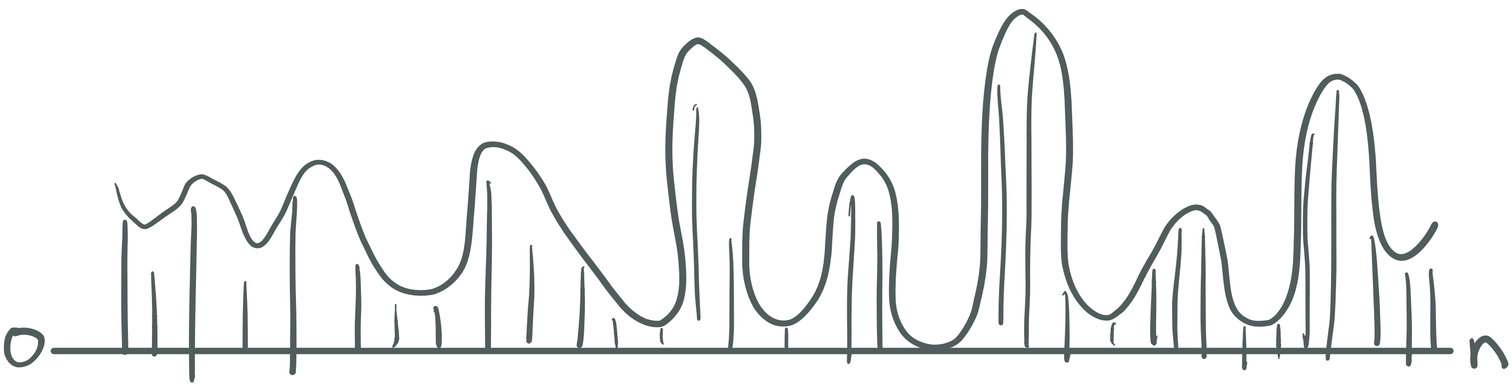
Goal: return the median w/ minimum:

(a) # passes (b) space

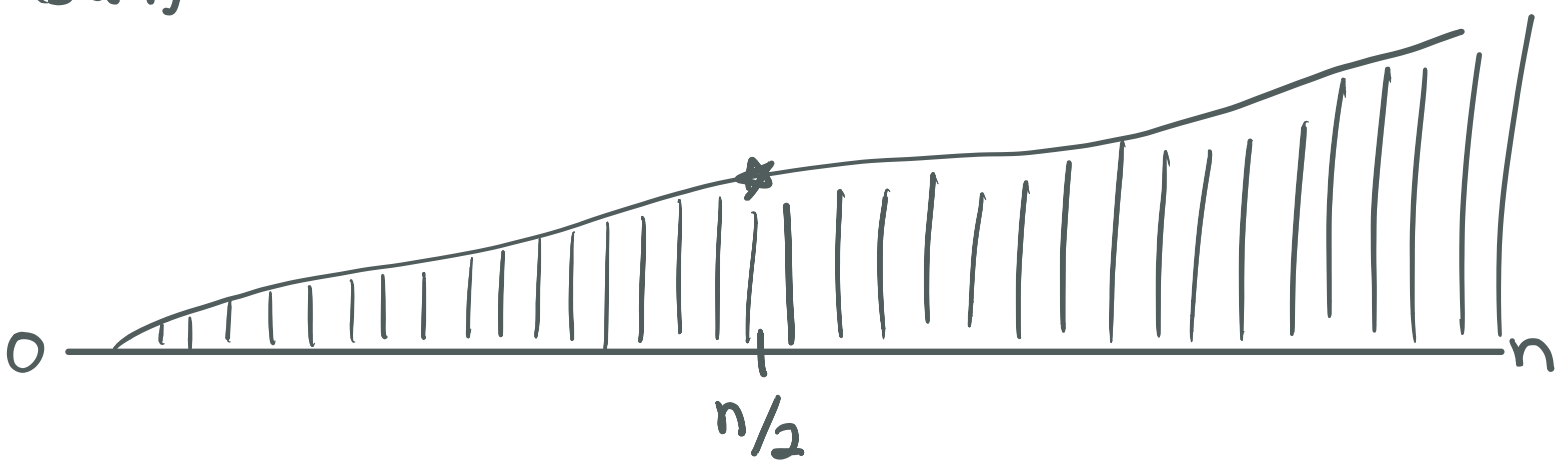
more generally: "quantile queries"

select rank k element
(kth largest)

1 pass: Input



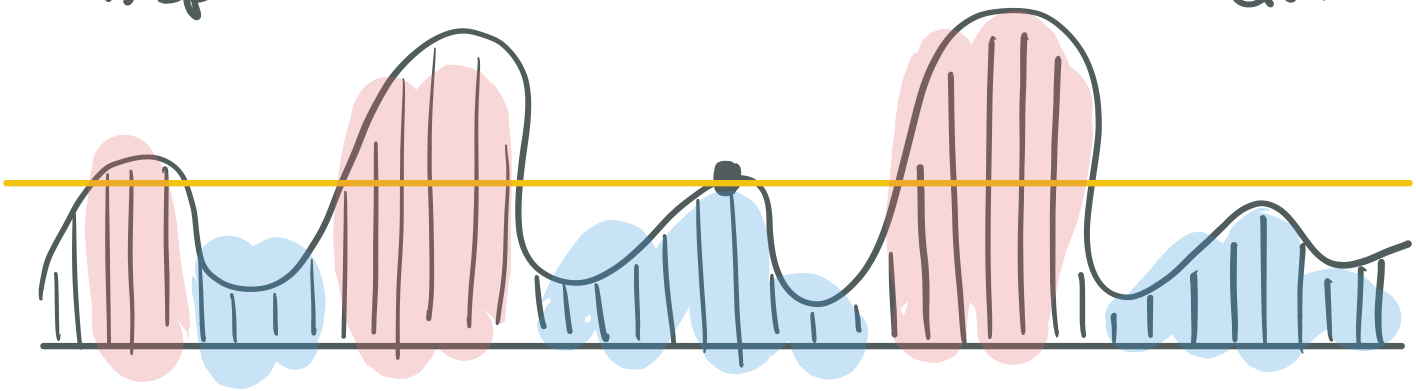
(sort)



sort and select

O(1) space

Quickselect



count # above, below pivot

recurse on appropriate half

Passes	Space	
1	$O(n)$	sort and select
$O(\log n)$	$O(1)$	quickselect (random pivot)
p	$\tilde{O}(n^{1/p})$	Munro, Paterson [1980]

this lecture

Approximations

given rank $k \in [n]$ and param $\varepsilon > 0$,

return element w/ rank $k \pm \varepsilon n$

Sampling:

for median:

sample $\ell = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$ elements

return median of sample

for rank $k = \alpha n$

return rank $\alpha \ell$ element of sample

Deterministic?

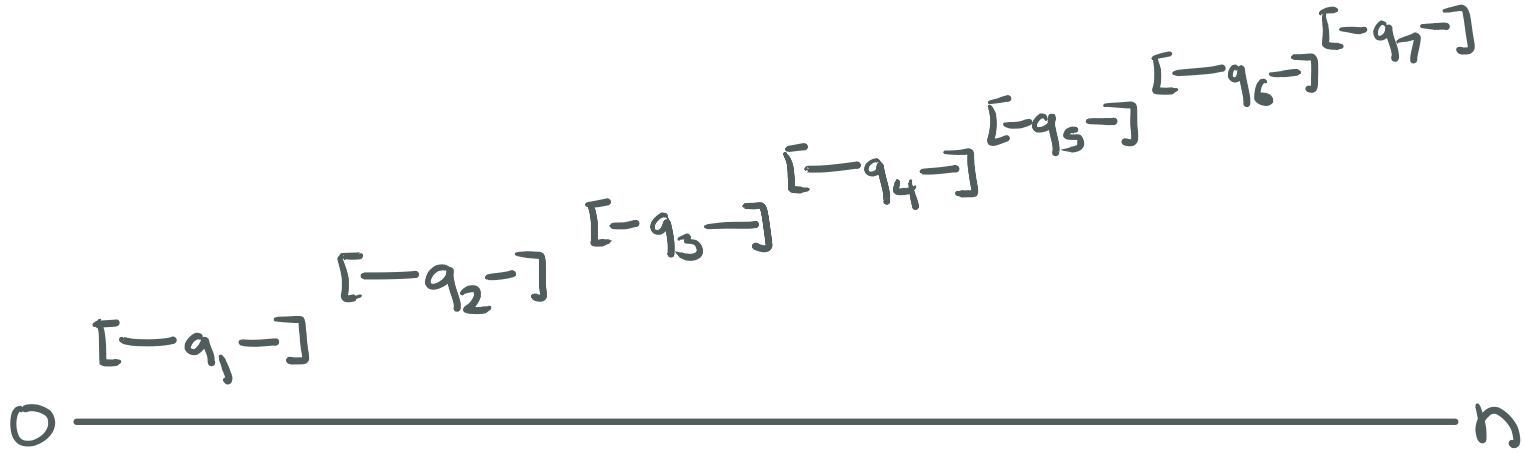
Quantiles

space-efficient
mergable
answer ε-approximate quantile queries

ℓ elements $q_1 < q_2 < \cdots < q_\ell$

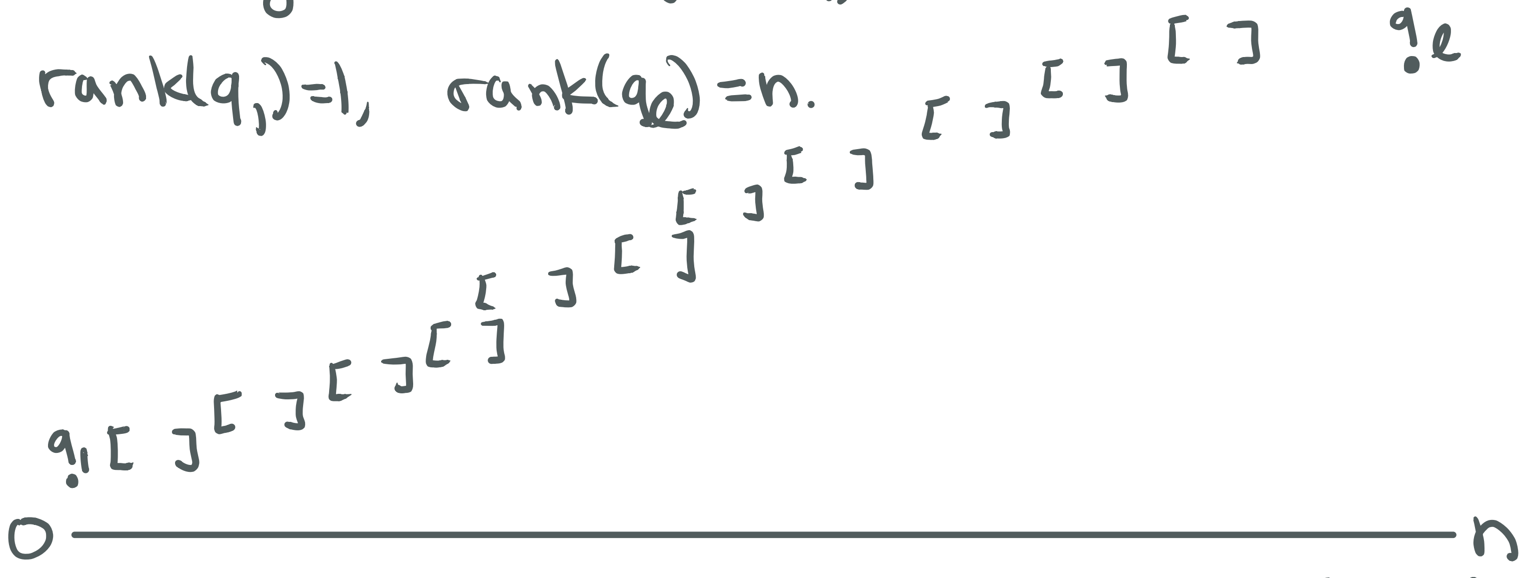
along w/ interval $I(q_i) \subseteq [1, n]$ s.t.

$\text{rank}(q_i) \in I(q_i)$

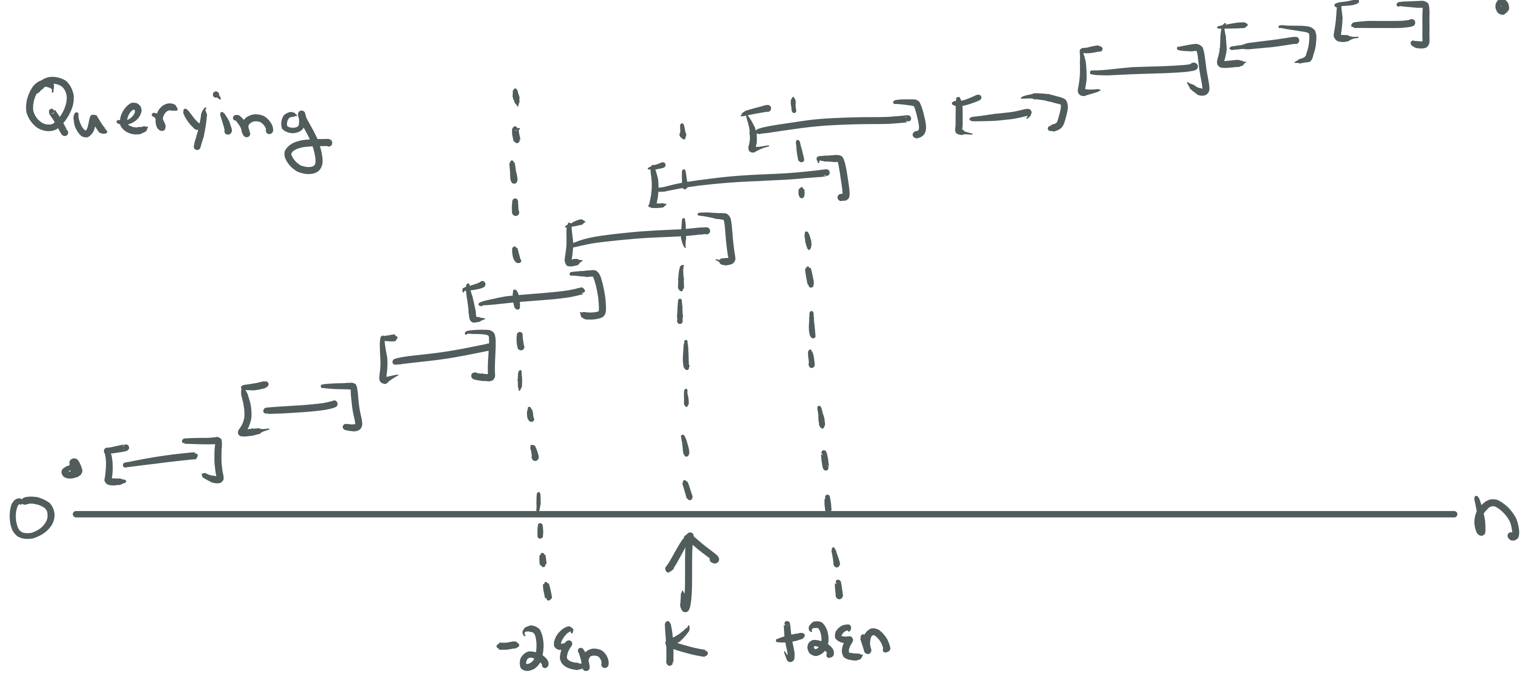


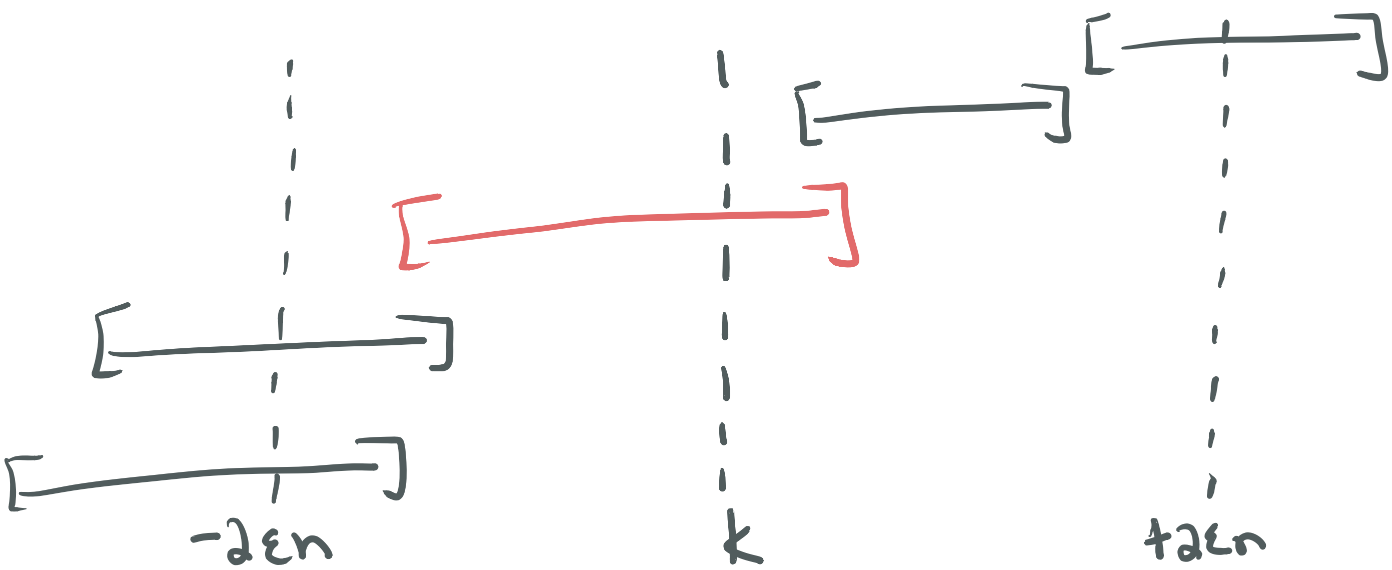
By tracking min/max specially, assume

$\text{rank}(q_1) = 1$, $\text{rank}(q_\ell) = n$.



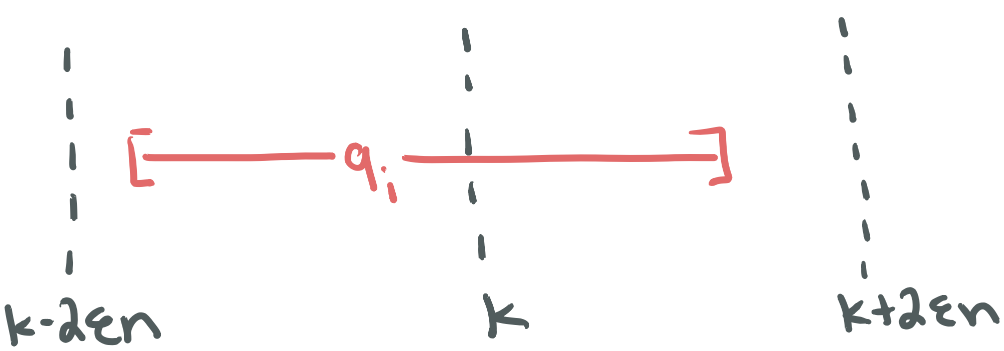
Querying





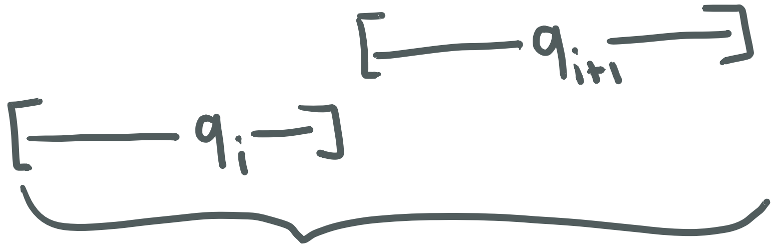
If $I(q_i) \subseteq [k-2\varepsilon n, k+2\varepsilon n]$ for some q_i,

then return q_i.



how to ensure such q_i exists $\forall k$?

lemma



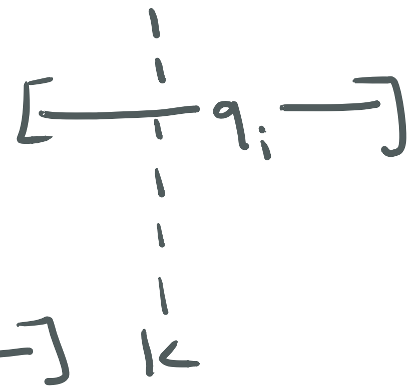
$\forall i$

total width $\le 2\varepsilon n$

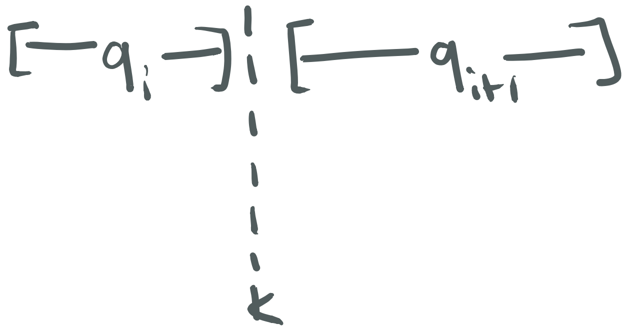
<u>Then</u> every query k contains an interval $I(q_i)$.

<u>Proof</u> two cases:

$k \in I(q_i)$ for some q_i



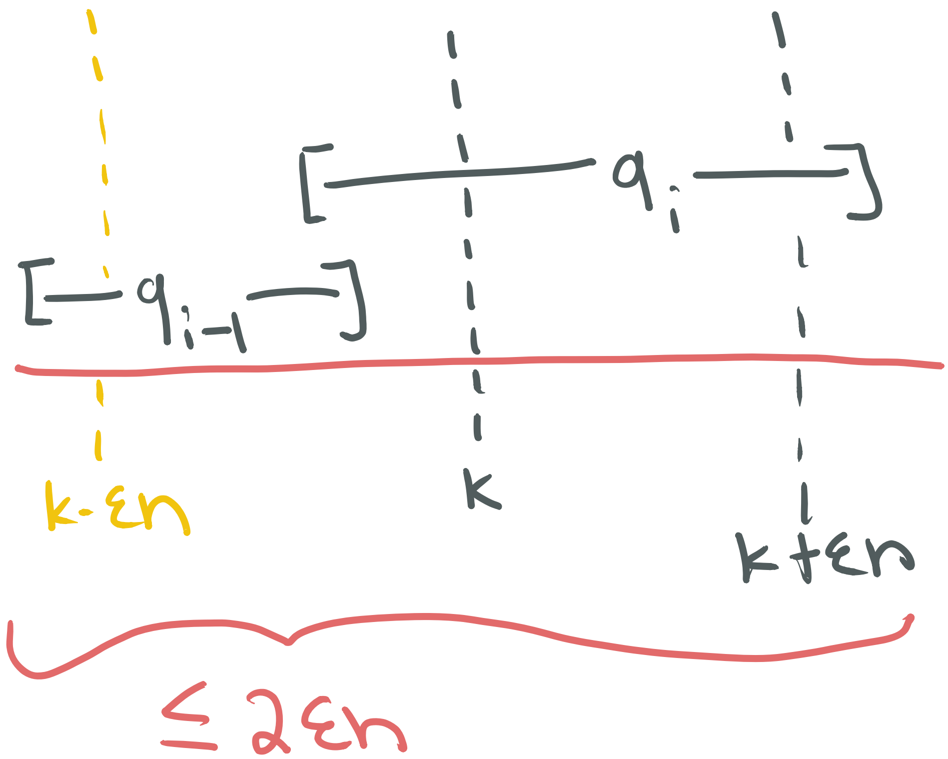
$k \notin I(q_i)$ $\forall q_i$



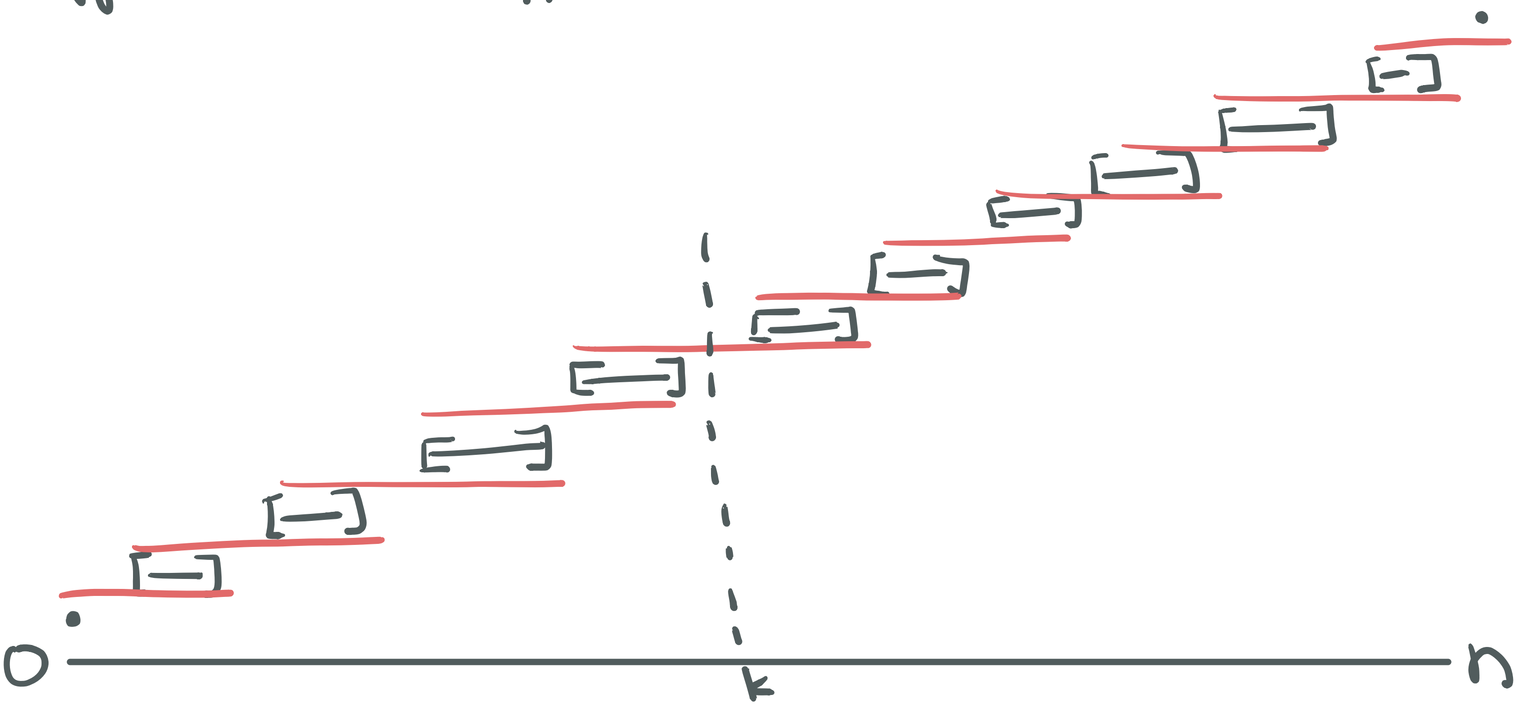
Suppose $k \in I(q_i)$ for some i

if $I(q_i) \subseteq [k-2\varepsilon, k+2\varepsilon]$ then done

else look at q_{i-1}

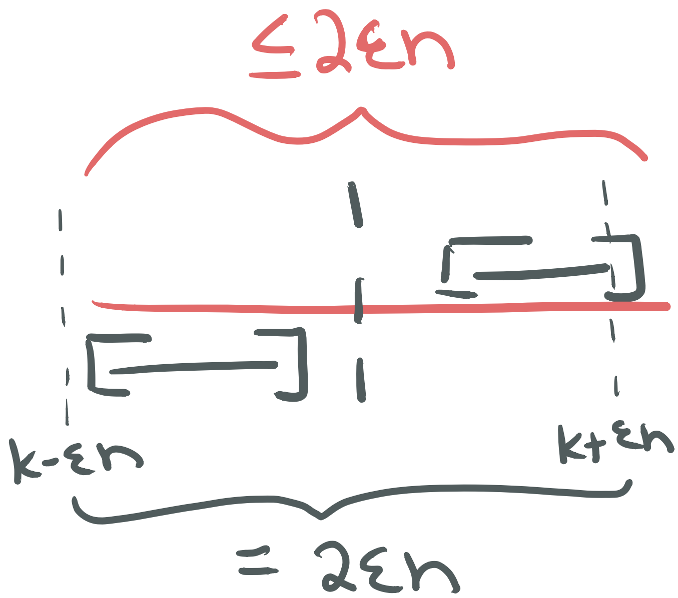


suppose $k \notin I(q_i)\ \forall i$



the "combined intervals" cover $[n]$

pick 1 covering k.



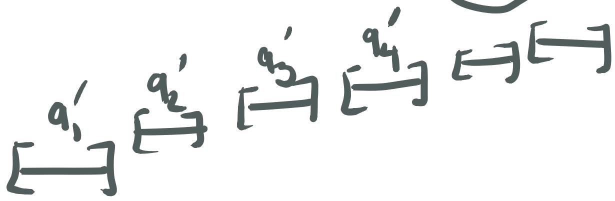
one of the intervals must lie inside

Key invariant: any two consecutive intervals have width $\leq 2\varepsilon n$.

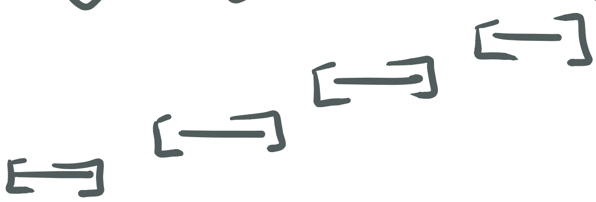
"ε-APX quantile summary"

<u>Merging</u> given two ε-APX quantile summaries over 2 streams, want ε-APX summary over <u>combined stream</u>

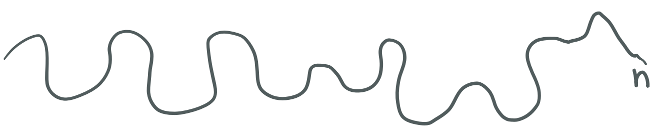
S_1  n_1

$\Rightarrow Q' =$ 

S_2  n_2

$\Rightarrow Q'' =$ 

want to combine Q', Q'' to get summary of

S_1  n_1 $+$ S_2  n_2

$$\text{"}Q' \cup Q''\text{"} = \{ q_1''', \dots, q_c''', I'''(q_1'''), \dots, I'''(q_c''') \}$$

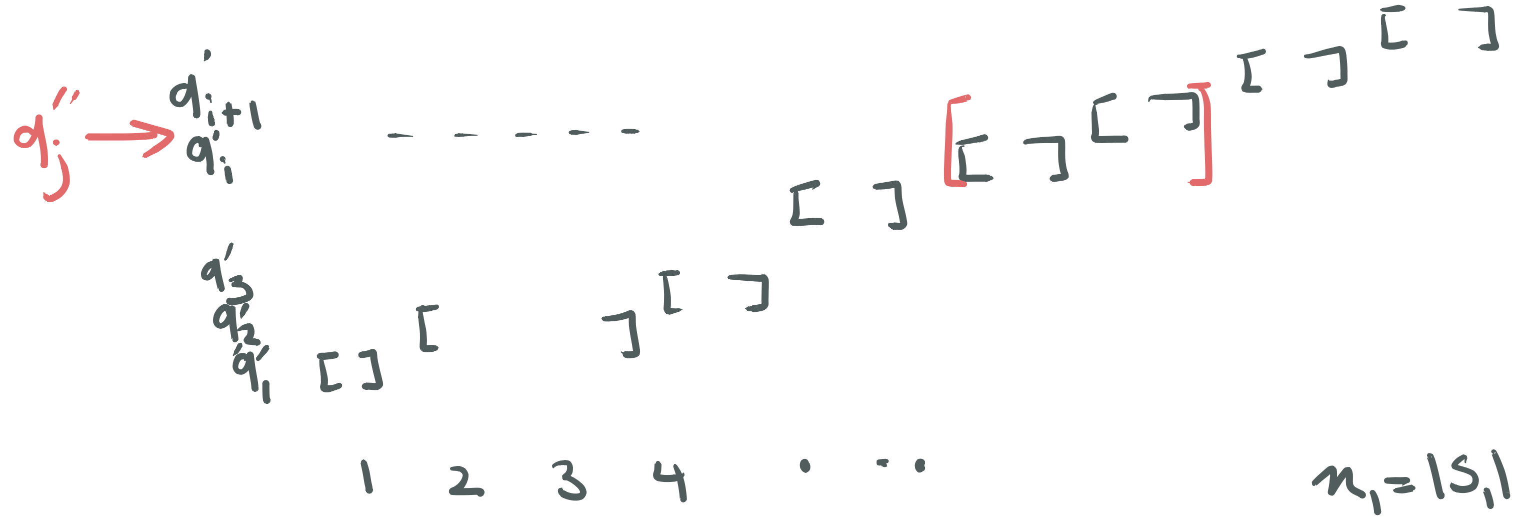
denote:

$$Q' = \{q'_1, \dots, q'_\ell, I'(q'_1), \dots, I'(q'_\ell)\}$$

$$Q'' = \{q''_1, \dots, q''_m, I''(q''_1), \dots, I''(q''_2)$$

let $q''_j \in Q''$. $I''(q_j)$ bounds rank q''_j w/r/t S_2

goal: bound rank q''_j w/r/t $S_1 + S_2$.



rank of q'_j in $S_1 \geq \min I'(q_i)$

$\leq \max I'(q_{i+1})$

so $\min I''(q'_i) + \min I''(q''_j)$

$\leq \text{rank}(q''_j \mid S_1 + S_2)$

$\leq \max I''(q''_j) + \max I''(q'')$

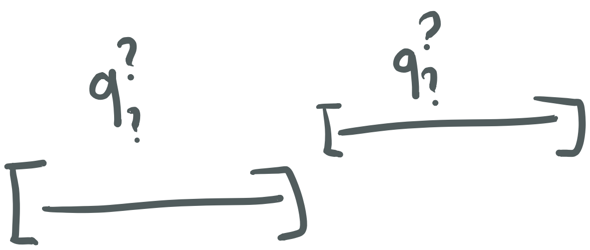
set $I'''(q''_j) = [\min I'(q'_i) + \min I''(q''_j),$

$\max I''(q'_{i+1}) + \max I''(q''_j)]$

$Q''' = \{q'_1, \ldots, q'_\ell, q''_1, \ldots, q''_m$, w/ intervals $I'''\}$

to show Q''' is ε-APX, need to show "$2\varepsilon n$ width" property.

Take two consecutive intervals in Q'''_3.



Two cases:

1. elements from diff sets
2. elements from same sets

① diff sets:

$[\text{---}\, q'_{i+1} \,\text{---}$

$[\text{---}\, q''_j \,\text{---}]$

$[\text{---}\, q'_i \,\text{---}]$

$\text{---}\, q''_{j-1} \,\text{---}]$

$$\underbrace{\min(I'(q_i)) + \min(I''(q_{j-1}))}_{(**)}$$

$$\underbrace{\max(I''(q_j)) + \max(I'(q'_{i+1}))}_{(*)}$$

$$\begin{aligned}(*) - (**) = \; & \max(I'(q_{i+1})) - \min I'(q_i) \; \} \leq 2\varepsilon n_1 \\ + \; & \max(I''(q_j)) - \min I''(q_{j-1}) \; \} \leq 2\varepsilon n_2 \\ \hline & \leq 2\varepsilon(n_1 + n_2)\end{aligned}$$

$\blacksquare$

② same sets

$[\!-\!-\; q''_{j+1} \;-\!-$

$[\!-\!-\; q'_{i+1} \;-\!-]$

$[\!-\!-\; q'_i \;-\!-]$

$q''_j \;-\!-]$

$\min I'(q'_i) + \min I''(q'_j)$

$\max I'(q'_{i+1}) + I''(q''_{j+1})$

$$
\begin{array}{r}
\max I'(q'_{i+1}) + I''(q''_{j+1}) \\
-\quad \min I'(q'_i) + \min I''(q'_j) \\
\hline
\varepsilon n_1 \quad + \quad \varepsilon n_2 \quad = \quad \varepsilon(n_1+n_2)
\end{array}
$$

$\blacksquare$

This shows that merging 2 ε-APX QS's gives ε-APX QS of combined streams

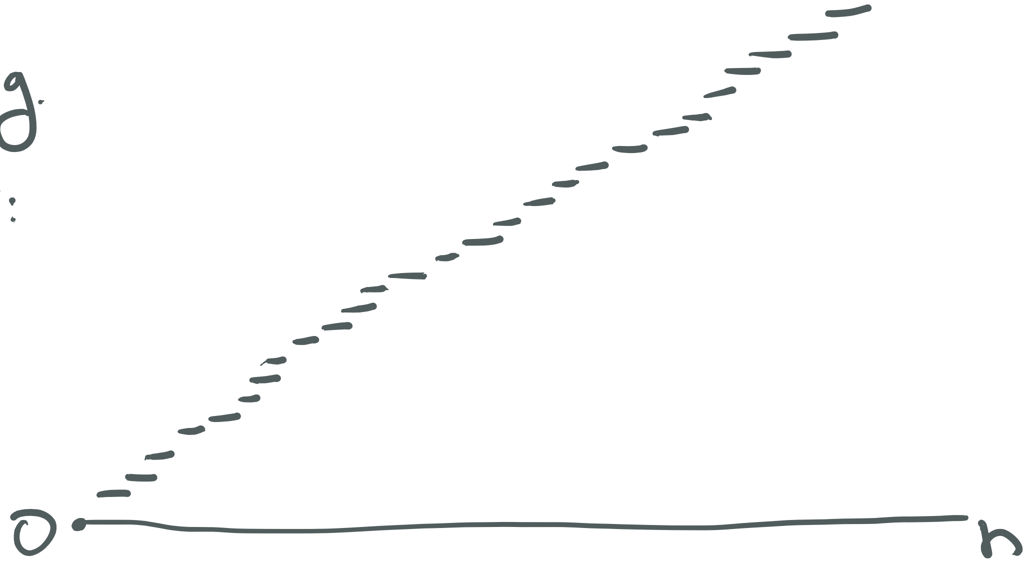
Lemma

given ε-approximate quantile summaries $Q_1, \ldots, Q_h$ for $S_1, \ldots, S_h$, we can combine to make ε-approximate summary $Q_1 \cup \cdots \cup Q_h$ of $S_1 \cup \cdots \cup S_h$.

Size?

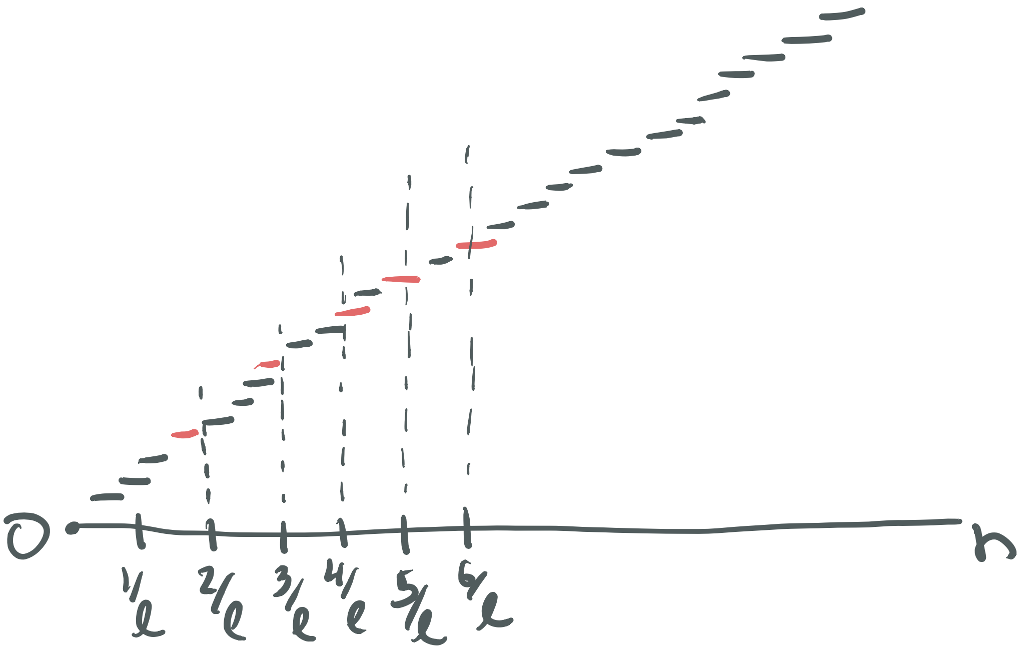
Pruning

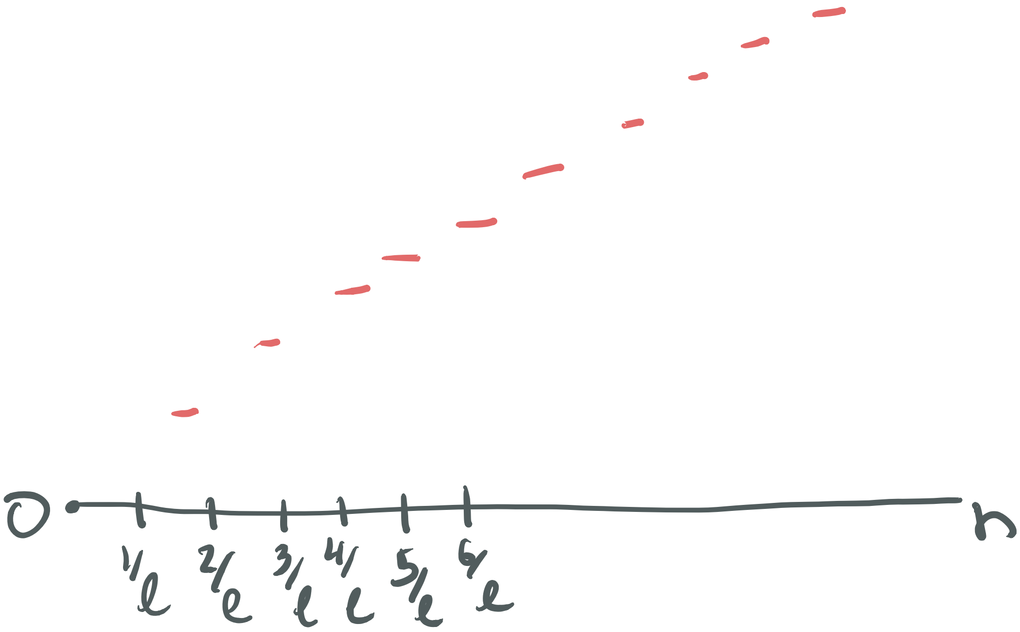
Input:



ε-approximate quantile summary w/ too many points

Goal: sparser summary that's still very good



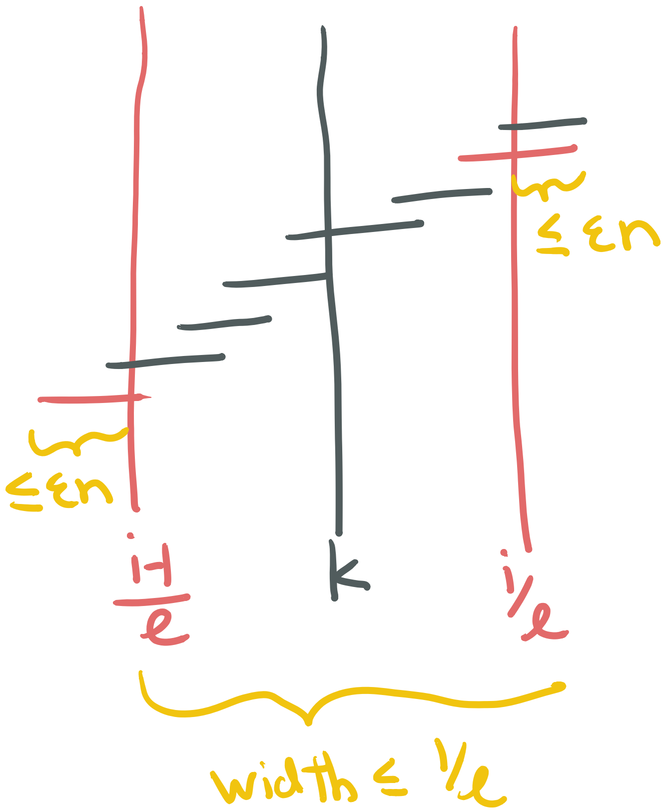


claim: resulting quantile is $(\varepsilon + \frac{1}{2\ell})$-APX

Proof

suppose we query a particular rank k

look at original summary



total width ≤ $2\varepsilon n + 1/\ell$

i.e. $(\varepsilon + \frac{1}{2\ell})$-approx.

Recap:

we can combine ε-APX quantile summaries to get ε-APX quantile summary of whole thing

sparsify ε-APX quantile summary to $(\varepsilon+\frac{1}{2k})$-APX quantile summary w/ k points

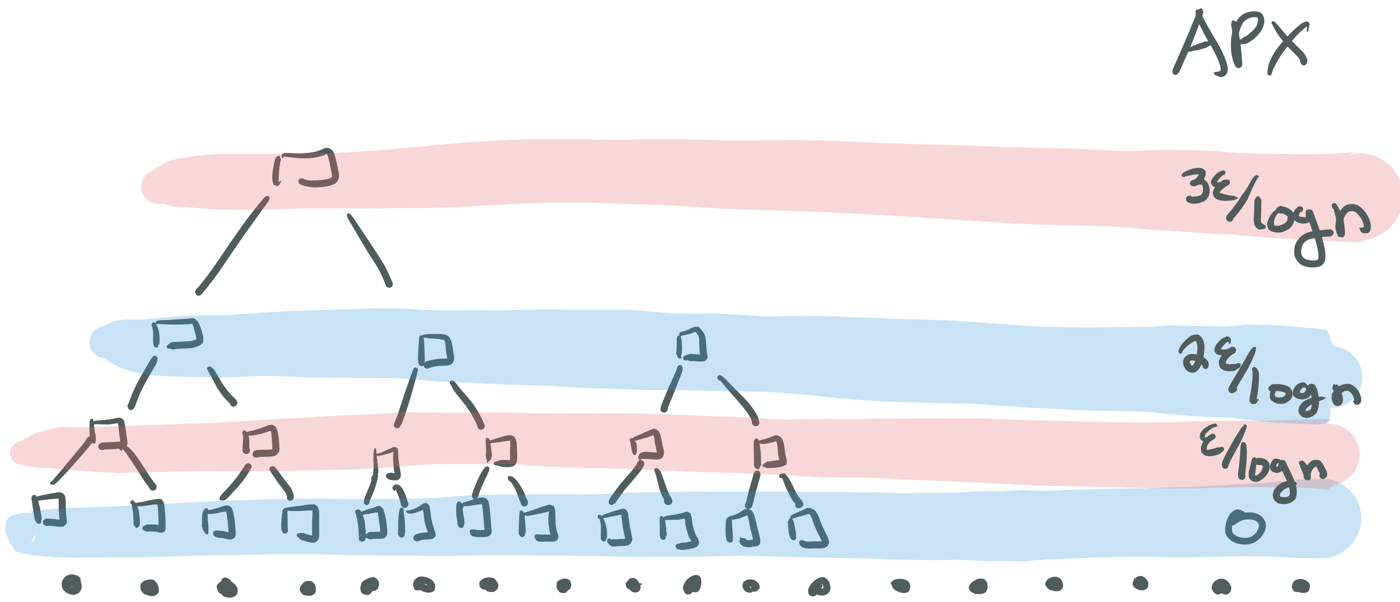
Remains to address:

how to make one at all??

what if $n=1$?

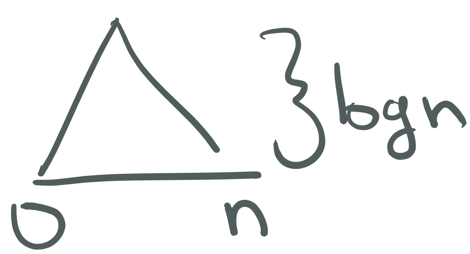
make a summary of just that point.

I claim that's all we need!



take $k = \frac{\log n}{\varepsilon}$

at the root,

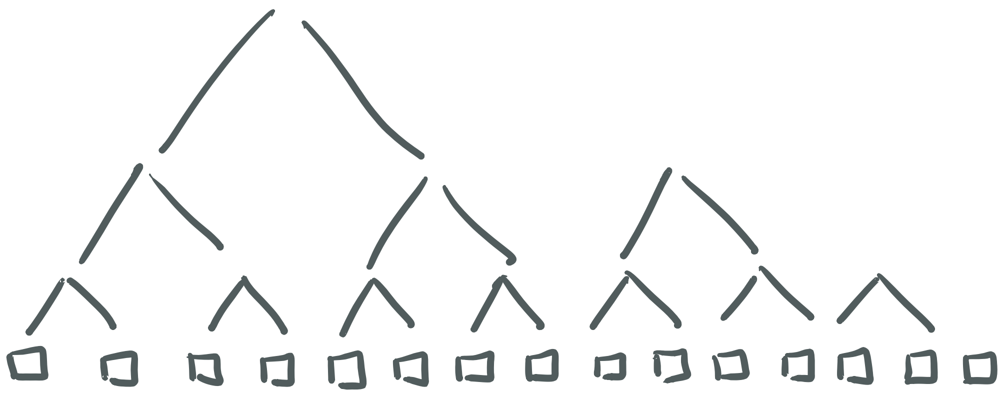


ε-approximate quantiles

Space?

each summary takes $O(\log(n)/\varepsilon)$ space.

How many summaries?



only keep "root summaries"

how many "root summaries"?

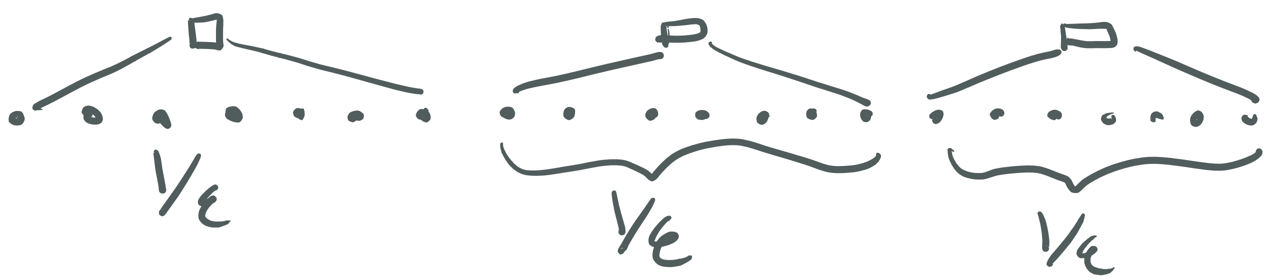
One summary per height, so $\log n$ total summaries at any time

Theorem

- 1-pass
- $O(\log^2(n)/\varepsilon)$ space
- deterministic
- ε-APX quantile over stream

idea: mergability + dyadic intervals trick

slightly better:



have first level contain $1/\varepsilon$ points

εn summaries at "leaf level"

$\Rightarrow \log(\varepsilon n)$ height

Theorem ††

- 1-pass
- $O(\log^2(\varepsilon n)/\varepsilon)$ space
- deterministic
- ε-APX quantile over stream

Even better?

Khanna-Greenwald:

$\frac{1}{\varepsilon}\log(\varepsilon n)$ space

- more sophisticated quantile summary, merging
- interval trick

Finding the median (and other ranks) in p passes

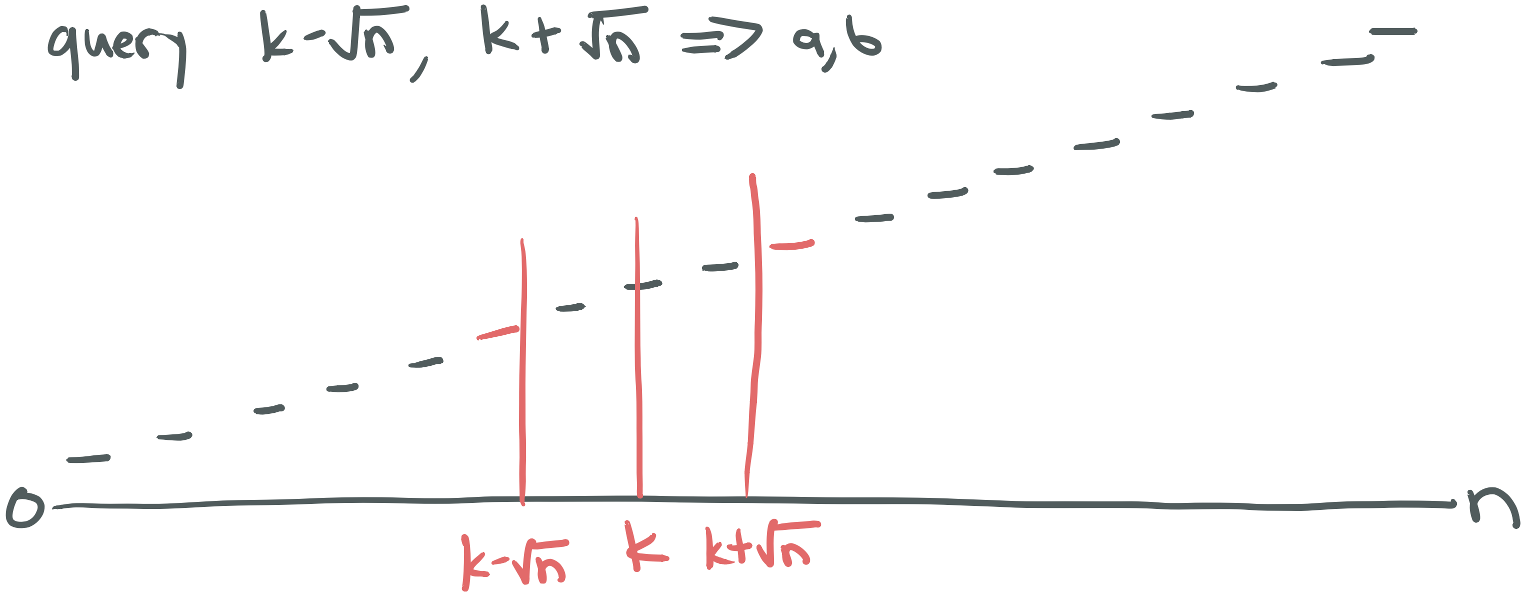
Fix $p=2$ for simplicity.

goal: $O(\sqrt{n}\ \mathrm{polylog}(n))$ space

suppose we are querying rank k.

1st pass: build ε-APX quantile summary for $\varepsilon = 1/\sqrt{n}$ ($\sqrt{n}\log(n)$ space w/ GK)

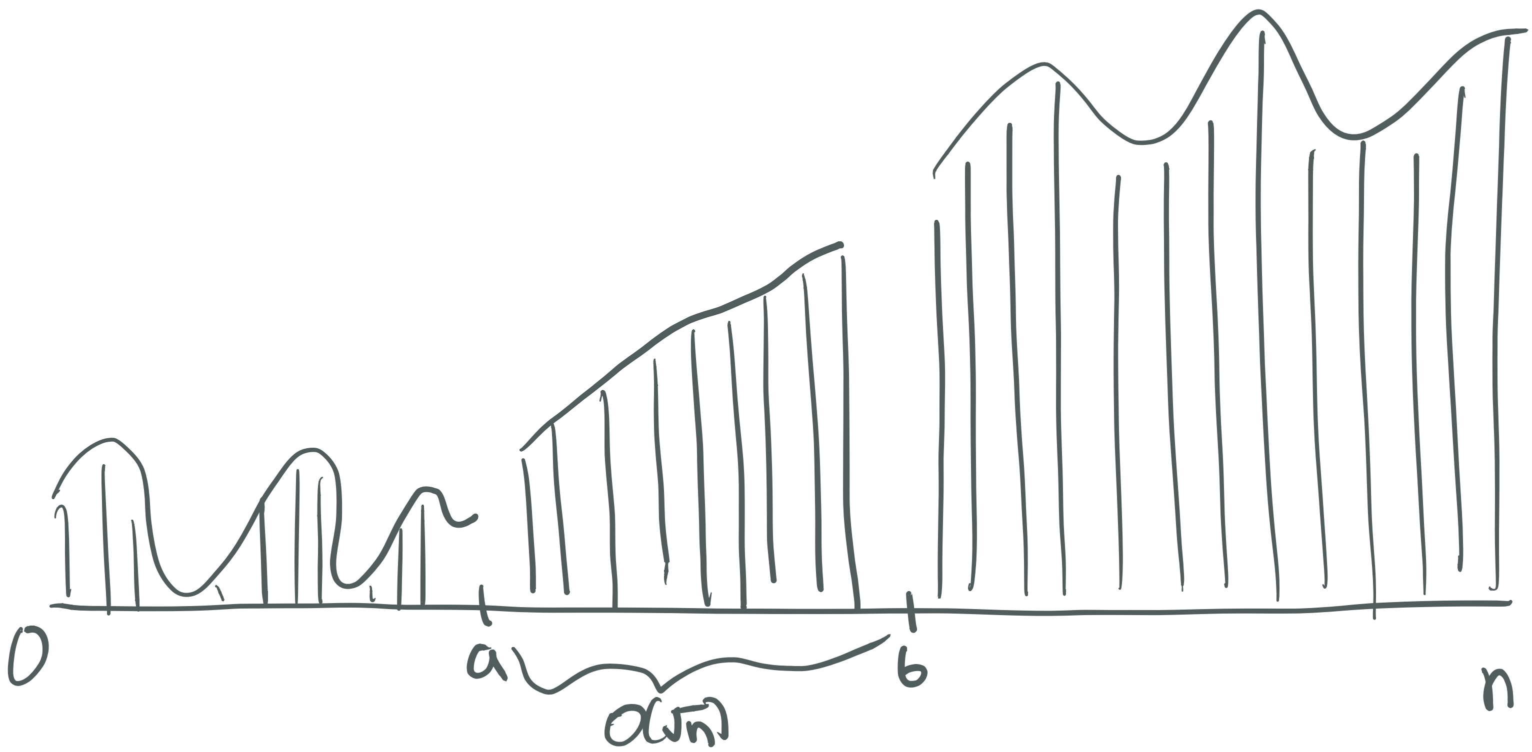
query $k-\sqrt{n}$, $k+\sqrt{n}$ $\Rightarrow$ a, b



$$k-2\sqrt{n} \le \mathrm{rank}(a) \le k \le \mathrm{rank}(b) \le k+2\sqrt{n}$$

2nd pass:

- take all elements between a and b



- compute rank(a), (by counting)
- sort taken elements
- return $(k-\text{rank}(a))$th taken element

for general p:

make $\frac{1}{n^{1/p}}$-APX quantile summaries and filter.

After p passes, down to $n^{1/p}$ elements sort and select.